Gaia: At the frontiers of astrometry
Editors : will be set by the publisher
EAS Publications Series, Vol. ?, 2010

TELESCOPES DON'T MAKE CATALOGUES!

David W. Hogg[1] and Dustin Lang[2]

Abstract. Astronomical instruments make intensity measurements; any precise astronomical experiment ought to involve modeling those measurements. People make catalogues, but because a catalogue requires hard decisions about calibration and detection, no catalogue can contain all of the information in the raw pixels relevant to most scientific investigations. Here we advocate making catalogue-like data outputs that permit investigators to test hypotheses with almost the power of the original image pixels. The key is to provide users with approximations to likelihood tests against the raw image pixels. We advocate three options, in order of increasing difficulty: The first is to *define* catalogue entries and associated uncertainties such that the catalogue contains the parameters of an approximate description of the image-level likelihood function. The second is to produce a K-catalogue *sampling* in "catalogue space" that samples a posterior probability distribution of catalogues given the data. The third is to expose a web service or equivalent that can *re-compute on demand* the full image-level likelihood for any user-supplied catalogue.

In probabilistic inference, the goal is to transform information gathered by the data-taking device into information about the model or parameters of interest, with as little loss as possible. The most precise methods for inference involve "forward modeling" of the measurements: A model that can generate the data accurately is a good model, and is constrained by every data element that it (interestingly) generates. For this reason, everyone who wants to perform a precise experiment with a telescope wants to model the image pixels. Recently, this has been realised in a range of astrophysics domains, from weak lensing (Bernstein & Jarvis 2002) to astrometry (Anderson *et al.* 2008; Lang *et al.* 2009).

Unfortunately, since its earliest days, astronomy has been done through *catalogues*. Historical data sets were made by eye, but enormous catalogues such as *USNO-B* (Monet *et al.* 2003), *2MASS* (Skrutskie *et al.* 2006), and *SDSS*

[1] Center for Cosmology and Particle Physics, New York University *and* Max-Planck-Institut für Astronomie, Heidelberg
[2] Princeton University Observatory

DOI: (will be inserted later)

(Abazajian *et al.* 2009) are based on digital intensity measurements. Despite this—for practical purposes—the primary data products of those surveys are catalogues; the survey teams considered it unimportant to give easy world-wide access to all the raw imaging pixels.

If an investigator wants to know the J-band fluxes of a few million *SDSS* sources using *2MASS* (something we often want to know), the only easy option is *catalogue matching* in which the investigator makes up heuristic matching conditions, tests them on a subset of sources, and then runs them on the full catalogues. For sources that don't get matched, what has the investigator learned? Very little, because it is almost impossible without access to the imaging pixels to determine the sensitivity of the *2MASS* data to the source at that source position, and even if the sensitivity is nominal, a two-sigma measurement (or even zero-sigma) is much more constraining than the uninformative statement that the source did not satisfy the *2MASS* catalogue-inclusion criteria. Furthermore, even when sources *do* match there are fundamental issues about what can be inferred (Budavári & Szalay 2008; Hogg & Lang 2008).

Most astronomical projects can be phrased as hypothesis tests. Thinking forward to *Gaia*, one important set of projects will involve finding streams of stars—linear features in phase space—that are likely remnants of disrupted stellar clusters or galaxies (*e.g.*, Helmi this volume). The detection of these lines of stars can be cast as a hypothesis test: If we put these stars onto a family of similar orbits, does our explanation of the data improve? Another set of projects will involve refinement of Milky Way parameters; these involve optimisation of some kind of likelihood, possibly marginalising out the details of the distribution functions (*e.g.*, Bovy *et al.* 2010a). Another set will involve testing the physical properties of the velocity-space structure in the Milky Way disk (Dehnen 1998; De Simone *et al.* 2004; Bovy & Hogg 2010b); these make different predictions for the *Gaia* data. The question of which is best is—in part—a question about their relative likelihoods under the data.

This article is about astronomical imaging in general and the data products derived therefrom, but the specific examples will be *Gaia*-related.

1 A definition of uncertainty

The simplest option for transmitting image information to the catalogue is to build the catalogue such that hypothesis tests against it are identical to—or as close as possible to—hypothesis tests against the raw imaging. Not only is this possible, it is practical in some cases.

For example, the current plan for the spectroscopic output of *SDSS-III BOSS* is that each individual *BOSS* spectrum will be a list of wavelengths λ_i and at each of those wavelengths a flux value f_i and associated inverse uncertainty variance $1/\sigma_i^2$. These fluxes and inverse variances will not be simply "best-fit" values; they will be the parameters of a model of the likelihood function (Bolton & Schlegel 2010): Imagine that for this spectrum an investigator has two possible models $m_1(\lambda)$ and $m_2(\lambda)$. The fluxes f_i and inverse variances $1/\sigma_i^2$ are constructed such that the

natural χ^2 difference

$$\Delta\chi^2 \equiv \sum_i \left[\frac{m_2(\lambda_i) - f_i}{\sigma_i^2}\right] - \sum_i \left[\frac{m_1(\lambda_i) - f_i}{\sigma_i^2}\right] \tag{1.1}$$

is as close as possible to the χ^2 difference the investigator *would have* obtained had he or she performed the χ^2 test in the raw spectrograph image pixels. Since χ^2 is related to likelihood by

$$\Delta\chi^2 \approx -2\,\ln\left[\frac{\mathscr{L}_2}{\mathscr{L}_1}\right] \quad , \tag{1.2}$$

this makes the f_i and $1/\sigma_i^2$ the parameters of an approximation to the pixel-level likelihood function. It is an approximation because it assumes Gaussian uncertainty, and small wavelength-to-wavelength covariance (this latter point is addressed in detail by Bolton & Schlegel 2010).

This suggests a general *definition* for catalogue elements and their associated uncertainties, which we explicitly advocate here. In the imagined case of *Gaia*, the definition would be this: For each star on the sky, there will be six (or more) parameters, say

$$\boldsymbol{y}^{\mathsf{T}} = [\mathrm{RA}, \mathrm{Dec}, \varpi, \mu_\alpha, \mu_\delta, v_r] \tag{1.3}$$

and an associated six-by-six inverse covariance matrix $\boldsymbol{C}^{-1}$. Imagine that for this star an investigator has two proposals $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ about the six parameters. We recommend that the catalogue entries $\boldsymbol{y}$ and $\boldsymbol{C}^{-1}$ be *defined* such that the natural χ^2 difference

$$\Delta\chi^2 \equiv [\boldsymbol{Y}_2 - \boldsymbol{y}]^{\mathsf{T}} \cdot \boldsymbol{C}^{-1} \cdot [\boldsymbol{Y}_2 - \boldsymbol{y}] - [\boldsymbol{Y}_1 - \boldsymbol{y}]^{\mathsf{T}} \cdot \boldsymbol{C}^{-1} \cdot [\boldsymbol{Y}_1 - \boldsymbol{y}] \tag{1.4}$$

is *as close as possible* to the logarithmic likelihood ratio $-2\,\ln[\mathscr{L}_2/\mathscr{L}_1]$, now *marginalised over all calibration and instrument parameters*, that you would have measured if you had access to the raw image pixels and racks of metal. The proposal is to adopt this definition; adoption of this is essentially equivalent to obtaining the catalog entries and uncertainties from Gaussian fits to the marginalized likelihood.

The nuisance-parameter marginalization takes this beyond the *BOSS* plan and makes it challenging. Marginalisation of a likelihood requires—in the mathematical sense—a prior probability distribution function (PDF) over the parameters to be marginalised out and it looks something like this:

$$\begin{aligned} \mathscr{L}(\boldsymbol{Y}) &\equiv p(\boldsymbol{D}|\boldsymbol{Y}) \\ &= \int \mathrm{d}\boldsymbol{\theta}\, p(\boldsymbol{\theta})\, p(\boldsymbol{D}|\boldsymbol{Y}, \boldsymbol{\theta}) \quad , \end{aligned} \tag{1.5}$$

where the object $\boldsymbol{D}$ is the entire raw data set (pixel intensities), $\boldsymbol{Y}$ is the set of six parameters of the individual object, and $\boldsymbol{\theta}$ is the set of all calibration parameters in the model, including all those related to attitude and hardware. The unmarginalised likelihood—the probability of the data given the star and calibration

parameters—depends, of course, on both the star and the calibration parameters; the marginalisation is the only conservative and accurate way to propagate uncertainties about calibration into the astrophysical parameters of interest. The prior $p(\boldsymbol{\theta})$ is required as a measure on the parameter space that permits integration.

If the entries are defined this way, a catalogue user can make approximations to likelihood comparisons at the pixel level. Unfortunately, this proposal is only sensible when the marginalized likelihoods are close to Gaussian in form and when star–star covariances can be ignored. These covariances are not expected to be negligible for *Gaia* (see Holl this volume). However, this will be a problem for almost all of the currently conceived plans for the *Gaia* catalogue. Transmission of star–star covariances is permitted by our next proposal.

2 A sampling in catalogue space

At the present day, in most situations of fitting large (non-parametric, or highly parameterised) models to large data sets, investigators report uncertainty information through *samplings*. A sampling has the disadvantage that it is a mixture-of-delta-functions approximation, but it has the great advantage that it can, in principle, describe an arbitrarly complicated PDF. There are also excellent sampling technologies available (*e.g.*, MacKay 2003).

A useful sampling of *Gaia* catalogues could have the following properties: There would be K catalogues that represent a sampling—perhaps not a Poisson sampling, but a sampling—from the posterior PDF in what you might call "catalogue space". Any experiment or measurement is performed on all K samples. The variance across the K outcomes of the K identical experiments is an estimate of the uncertainty on the results of the outcome on the primary catalogue. In this vision, the the sampling provides a rank-K approximation to the full-catalogue covariance matrix (which is billions by billions and non-sparse).

Importantly, in this vision, the sampling of catalogues should be a sampling not just in the space of the astrophysical parameters but also in the space of the attitude and calibration parameters (the nuisance parameters). This ensures that the variance over samples contains the propagated uncertainty from the nuisance parameters, and it doesn't add significantly to any of the technical challenges of producing the sampling.

Although there is an enormous literature on sampling methods for large problems, there is a useful hack that might be appropriate if practical sampling methods fail on a problem of this scale: Approximate samples can be made by inflating leave-one-out jackknife trials. The idea would be to cut the raw data stream into K equally informative (roughly speaking, equal-sized) disjoint subsamples, and construct the K complementary leave-one-out subsamples, in each of which one of the K disjoint subsamples is left out. All data analysis (from raw data to inference of all nuisance and astrophysical parameters through to catalogue generation) is performed on each of the K leave-one-out subsamples. The *difference* (in catalogue space) between the primary catalogue and each of the K leave-one-out catalogues can be amplified by a factor of (very close to) K to produce a sampling that has

the same full-catalogue variance and covariance as the classical jackknife estimate. This might be an expedient way to produce and publish the jackknife estimate of the full error covariance, that has most of the good properties of a sampling.

Technically, a sampling is a description not of the likelihood but of a posterior PDF, and it has the disadvantage to its users that it has had a prior PDF multiplied in. This is a fundamental limitation of the approach, but for most kinds of catalogue outputs astronomers are interested in, the data are so informative that the posterior PDF in catalogue space looks very much like the likelihood function unless the adopted priors are extremely informative. Another issue is that additional information (say, from another telescope or instrument), which is expressed as a multiplicative likelihood term, can cause huge changes in the relative posterior probabilities of the K catalogue samples. Indeed, all K samples may be essentially ruled out by new information, or the samples may fail to capture an important mode in catalogue space. This is a fundamental limitation of sample-based representations.

In an ideal but challenging world, a sampling in catalogue space would explore regions of differing *complexity*. Not sure if a star is binary? Some of the K catalogues would have it binary and some would have it single. There is no reason in principle that the catalogues would be qualitatively similar in cases in which there is real statistical scientific uncertainty. This complexifies substantially sampling strategies, and complexifies downstream analyses by users, so we leave it here only as a comment. Changes to catalogue complexity are handled very naturally by our third proposal.

3 Exposing the full likelihood function

All this has been about approximations to the likelihood function, which begs the question: Why not just publish the likelihood function itself? Here what we imagine is an interface (perhaps in the cloud; see O'Mullane this volume) to which a user could submit a *catalogue diff*—a difference between the primary *Gaia* catalogue and the catalogue he or she wants to test. The machinery behind the interface would use the modified catalogue to generate the raw pixels, compare to the data, and return the *likelihood diff*. Of course the user must be permitted to modify also nuisance parameters, or else be presented with the option of obtaining the results marginalised over these.

An interface of this kind would be expensive to build, run, and maintain; it is probably impractical at the present day. However, it would permit literally arbitrary hypothesis testing, and arbitrary calculation of covariances. It would also make it the responsibility of users to perform their own uncertainty analysis and propagation, relieving the teams of some of their most burdensome duties. Extremely clever systems could cache computation so subsequent users could benefit from the users that have come before. However, without technology development, this proposal is probably impractical for the *Gaia* data.

4 Discussion

How big do you have to make K? There is no simple answer; this depends on the complexity of the posterior PDF and the precision with which users need to know their uncertainties. Our attitude is that you never—in any finite experiment—know your *uncertainties* to high precision, so we have the intuition that K does not have to be large to be an enormous improvement over a single catalogue with reported single-star uncertainties. One order-of-magnitude option for K in the case of *Gaia* is the number of *visits*, the number of times the spacecraft scans across a typical star. Jackknifing more finely than this will not help much on individual-star uncertainties.

But if anyone can make their own catalogue, the data will be changing; the science won't be repeatable! This point is good; published science needs to be repeatable by referees and subsequent investigators. On the other hand, our *understanding* of the *Gaia* data will necessarily evolve as calibrations and physical models evolve. There is no hope for a completely stable catalogue. What *is* fixed is the set of raw pixel data. The approaches advocated here make the raw data the primary objects of release, and therefore tie the release to the only part of the data analysis that *is* stable and repeatable. Of course, it is necessary for reasons of repeatability to clearly tag, cut, and release well-documented and stable versions of the data, likelihood function, and primary catalogue.

Is there any relevant difference between a Bayesian and a frequentist? Bayesianism is required for marginalization, because the prior provides the measure for integration. However, Bayesians and frequentists agree—or should agree—that the most important publication of an experiment is the *likelihood function*. It is through the likelihood that data impact our beliefs; different investigators have different priors, different objectives, and different external data at their disposal. They can only combine these with the *Gaia* data properly if the likelihood function is exposed. Even if the *Gaia* outputs end up being a sampling of some posterior PDF, those samples should come with evaluations of the prior, so that subsequent users can divide it out before combining with their own individual knowledge and data.

It is a pleasure to thank Anthony Brown, Francoise Crifo, Ken Freeman, and Lennart Lindegren, whose comments *during* the meeting shaped these ideas in real time, and Coryn Bailer-Jones, Jo Bovy, and Sam Roweis for valuable discussions. Partial financial support was provided by NASA (grant NNX08AJ48G), the NSF (grant AST-0908357), and a Research Fellowship of the Humboldt Foundation.

References

Abazajian, K. N. *et al.*, 2009, ApJS, 182, 543

Anderson, J. *et al.*, 2008, AJ, 135, 2114

Bernstein, G. M., & Jarvis, M., 2002, AJ, 123, 583

Bolton, A. S. & Schlegel, D. J., 2010, PASP, 122, 248

Bovy, J., Murray I., & Hogg, D. W., 2010a, AJ, 711, 1157

Bovy, J. & Hogg, D. W., 2010b, ApJ, 717, 617

Budavári, T., & Szalay, A. S., 2008, ApJ, 679, 301

Dehnen, W., 1998, AJ, 115, 2384

De Simone, R. S., Wu, X., & Tremaine, S., 2004, MNRAS, 350, 627

Hogg, D. W. & Lang, D., 2008, AIP Conf. Proc., 1082, 331 (arXiv:0810.3851)

Lang, D., Hogg, D. W., Jester, S., & Rix, H.-W., 2009, AJ, 137, 4400

Lang, D., Hogg, D. W., Mierle, K., Blanton, M., & Roweis, S., 2010, AJ, 139, 1782

MacKay, D., 2003, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press

Monet, D. G. *et al.*, 2003, AJ, 125, 984

Skrutskie, M. F. *et al.*, 2006, AJ, 131, 1163